GREQAM
Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes études en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail
n°2010-47

An Agnostic Look at Bayesian Statistics and Econometrics

Russell Davidson

November 2010

An Agnostic Look at Bayesian Statistics and Econometrics

by

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

russell.davidson@mcgill.ca

Abstract

Bayesians and non-Bayesians, often called frequentists, seem to be perpetually at loggerheads on fundamental questions of statistical inference. This paper takes as agnostic a stand as is possible for a practising frequentist, and tries to elicit a Bayesian answer to questions of interest to frequentists. The argument is based on my presentation at a debate organised by the Rimini Centre for Economic Analysis, between me as the frequentist "advocate", and Christian Robert on the Bayesian side.

Keywords: Bayesian methods, bootstrap, Bahadur-Savage result

JEL codes: C10, C11, C50

This research was supported by the Canada Research Chair program (Chair in Economics, McGill University), and by grants from the Social Sciences and Humanities Research Council of Canada and the Fonds Québécois de Recherche sur la Société et la Culture. This paper is based on my contribution to a debate that took place at the Rimini Centre for Economic Analysis in July 2009, between Christian Robert and me, entitled *The 21st Century Belongs to Bayes*. Special thanks go to Gael Martin for coordinating the debate, and nudging me to produce this paper in a reasonable amount of time.

February 2010

1. Introduction

Bayesians are of course their own worst enemies. They make non-Bayesians accuse them of religious fervour, and an unwillingness to see another point of view. I'm not going to engage in a religious war – history shows that such wars achieve nothing at a very high cost. In a sense, it is easy for non-Bayesians to take cheap potshots at Bayesians, and I hope I will resist the temptation to do that. Rather, I will expound some matters of current concern to frequentist econometricians, and ask some questions about what a Bayesian take on these might be. It is no doubt just as easy for Bayesians to take cheap potshots at unthinking frequentists, who may think, for instance, that the problems posed by some asymptotic theory or another are fundamental, while a Bayesian perspective reveals these to be non-problems. I won't take the time to defend such frequentists either.

The outlook I take in this paper is what I term "agnostic". By that I mean not only that I won't come down on one side or the other, for or against Bayesian methods, but further that I don't see any interest in trying to do so. It seems plain that both Bayesian and non-Bayesian approaches have contributed a lot to our understanding of the workings of the economy, and of important issues in many other disciplines.

Notwithstanding, in my presentation at the Rimini debate, I set myself up as the Counsel for the Prosecution, apparently making a case against Bayesian methods, but in fact trying to pose some hard questions about how a "practising" Bayesian would set about tackling issues that have earned much recent attention in the literature styled as "frequentist"; all of this from the point of view of a "practising" frequentist. In taking this position, I was aware that I threw Christian Robert, my "opponent", on the defensive, but this was not just a debating trick (although I confess it was that), but also a way for me to elicit answers from Bayesians to questions to which I genuinely had no answers. Since as a frequentist I was very much in the minority at the Rimini meeting, I convinced myself that my tactics were legitimate, and I believe that, both at the meeting in many of the presentations made there and in Christian's reply, satisfactory answers to many of my points were made.

2. What I won't discuss

Out of the many points that have traditionally aroused passions on one side or the other, here is a list of those that, as a good agnostic, I do not wish to discuss. Some interesting discussion of some of these points is found in an exchange between Gelman (2008) and Senn (2008), that has a certain resemblance to the current exchange between Christian Robert and me. I try not to repeat their discussion.

- The likelihood principle as fundamental to all inference.

I think that many, and probably most, Bayesians do adhere to this principle. Christian Robert asserts that he "simply does not believe meaningful inference is possible without this likelihood function." Well, I don't agree, but I won't attack this issue head on, as I think that there are quite deep issues, about which I will say more in the next section, concerning meaningful inference, about which we may agree independently of belief or not in the likelihood principle.

- Point estimates (with standard errors) are inferior/superior to a posterior distribution.

Here I adopt a perfectly agnostic point of view. Since we have all learned many useful things using each one of these approaches, it seems unnecessarily combative to try to rank them.

- The Bayesian approach is the only coherent method for incorporating prior knowledge.

There are many circumstances in which we really, honestly, do have prior knowledge, and I am happy to concede that the standard Bayesian way of combining a prior distribution with the information in a data set, as represented by a likelihood, is much more straightforward to implement that any non-Bayesian approach that I know of. This is a topic where frequentists, or other non-Bayesians, might well devote some research effort to try to come up with a comparably simple methodology that doesn't involve assigning a probability measure to the parameter space.

- Probability theory is mathematics, and so is not about anything until we apply it to something.

No fight from me about different applications of the same mathematical structure. One of the points of disagreement between Bayesians and frequentists is what is modelled as random, and what as deterministic. Above I allude to a probability measure on the parameter space. Why not? No mathematical reason, certainly, and so, if you can come up with a useful interpretation of such a measure, as Bayesians do, there's no reason to suppose that there's anything wrong with two different applications of the same mathematical theory, one Bayesian, the other frequentist.

- Frequentists make (too frequent?) use of asymptotic theory.

I have heard frequentists reproached by Bayesians for their belief in "asymptopia", that is, their regular use of asymptotic theory. I am no great friend myself of many of the asymptotic methods currently in use. But asymptotic theory is just a way of getting tractable approximations to distributions the exact forms of which are analytically intractable. As soon as something better comes along, like the bootstrap, of which more later, then I for one am happy to leave asymptotics behind.

- Bayesian methods are often too computer intensive.

Here I might take a potshot at those people on both sides of the fence who put no value on attempts to reduce the computational burden of various computer-intensive techniques, claiming that it's more effective to let the computer spend time on a problem for which a known solution method exists than to devote valuable research time to speeding up the computations. But research ostensibly aimed at improving algorithms has led many times to improved understanding of the procedures that make use of the algorithms. I feel strongly that it is important to *understand* what the computer is up to when it provides us with an "answer", and not just treat computer software as black boxes.

- Bayesian/frequentist methods are unscientific.

We are all trying to be as "scientific" as we can in our use of econometrics, and I see no reason for either side to accuse the other of being "unscientific". Unfortunately such accusations have been too frequent in both camps.

- There are many difficulties in formulating Bayes' Theorem in the formalism that is used, say, in financial economics/econometrics.

I'm not going to argue about its status as a mathematical result. It was nonetheless a surprise for me that the only formulation of Bayes' theorem I could find that sits on top of the usual $(\Omega, \mathcal{F}, P)$ definition of a probability space was singularly hard to interpret. I suspect that this is just another manifestation of the fact that the Bayesian and frequentist approaches are different.

3. Genuine difficulties for meaningful inference

For a frequentist, a **model** is a set $\mathbb{M}$ of data-generating processes (DGPs). Usually some structure is imposed on the set, the most important being the **parameter-defining mapping**, which maps $\mathbb{M}$ into a parameter space $\Theta \in \mathbb{R}^k$. It is often possible to extend the parametric structure to a group structure. There are so many ways to do this that I won't discuss them here: suffice it to say that many of the attempts to import the ideas of differential geometry into statistics make use of such a structure.

If the group is locally compact, then there exists a **left Haar measure** on the group, unique up to a multiplicative constant, which is invariant under left-translation by the group operation; see, for an early reference Halmos (1950), and a more recent expository one, Rubinstein-Salzedo (2004). This property makes the Haar measure, defined either on the parameter space or on the set $\mathbb{M}$, an ideal candidate for an uninformative prior.

However, frequentists, and, I suspect, many Bayesians, use models that are not locally compact in any usable topology. No Haar measure can be defined on these models, viewed as groups or not, and so no uninformative prior exists for such models. Here we can take note of the various attempts to define "improper" priors, noting at the same time that this has been a source of much controversy among Bayesians themselves.

There exists a much less well-known problem in frequentist theory, one that arises from similar topological considerations with models that are not locally compact. The problem is set out in a classic paper, Bahadur and Savage (1956), that has been scandalously neglected except by a few conscientious frequentists, see for instance Dufour (2003). Here, I wish to illustrate the problem they raise, because it strikingly sets limits to the extent to which frequentist inference is possible if one tries to make use of too unconstrained a model. If one can reasonably draw an analogy between the use of an informative prior by a Bayesian and a frequentist's use of a model with assumptions that restrict the sorts of probability distributions the model admits, then there is a message from Bahadur and Savage's result that is relevant to both camps, namely that no valid inference is possible unless prior information, in some form or another, is available.

The paper of Bahadur and Savage contains a number of impossibility results about inference on the expectation of a distribution based on an IID sample drawn from it. The thrust of all the results is that, unless some restrictions, over and above the mere existence of the expectation of the distribution, are placed on the class of distributions that constitute the model, such inference is impossible. Impossible in the sense that the size of a test of a specific value for the expectation is independent of the significance level, and that no valid confidence intervals exist.

The model for which these impossibility results hold must be reasonably general, and the precise regularity conditions made by Bahadur and Savage are as follows. Each DGP of the model is characterised by a cumulative distribution function (CDF), F say. The class $\mathcal{F}$ of those F that the model contains is such that

(i) For all $F \in \mathcal{F}$, $\mu_F \equiv \int_{-\infty}^{\infty} x\, dF(x)$ exists and is finite;

(ii) For every real number m, there is $F \in \mathcal{F}$ with $\mu_F = m$;

(iii) $\mathcal{F}$ is convex.

Let $\mathcal{F}_m$ be the subset of $\mathcal{F}$ for which $\mu_F = m$. Then Bahadur and Savage prove the following theorem.

Theorem 1

> For every bounded real-valued function ϕ defined on the sample space (that is, $\mathbb{R}^n$ for a sample of size n), the quantities $\inf_{F \in \mathcal{F}_m} \mathrm{E}\phi$ and $\sup_{F \in \mathcal{F}_m} \mathrm{E}\phi$ are independent of m.

From this theorem, the main results of their paper can be derived. The argument is based on the fact that the mapping from $\mathcal{F}$, endowed with the topology of weak convergence, to the real line, with the usual topology, that maps a CDF F to its expectation μ_F is not continuous. (Actually, Bahadur and Savage use a seemingly different topology, based on the metric of **absolute-variational distance**, defined for two CDFs F and G as

$$\delta(F, G) = \sup_{\phi \in \Phi} |\mathrm{E}_F \phi - \mathrm{E}_G \phi|,$$

where Φ is the set of real-valued functions of the sample space taking values in the interval $[0, 1]$.)

This failure of continuity is of course not specific to the first moment of a distribution. It applies equally well to all moments, and the impossibility result has recently been extended by Dufour to wider sets of circumstances, such as the linear regression model.

The Bahadur-Savage result casts a shadow of doubt over the whole enterprise of non-parametric statistics, where most theorists and practitioners want to make as few and as unrestrictive a set of assumptions as possible about the distributions of the random elements in their models. I now give an illustrative example of the problem, concrete enough for us to acquire some intuition about it. I present a one-parameter family of distributions, all with zero expectation. If an IID sample of size n is drawn from a distribution that is a member of this family, one can construct the usual t statistic for testing whether the

expectation of the distribution is zero. I will then show that, for `any` finite critical value, the probability that the t statistic exceeds that value tends to one as the parameter of the family tends to zero. It follows that, if all the DGPs of the sequence are included in $\mathcal{F}$, the t test has size one.

Each distribution in the family is characterised by a parameter $p \in [0, 1]$. A random variable from the distribution can be written as

$$U = Y/p^2 + (1 - Y)W - 1/p$$

where $W \sim \mathrm{N}(0, 1)$ and

$$Y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p, \end{cases}$$

W and Y being independent. It is evident that $\mathrm{E}_p U = 0$. Suppose we have a sample of n IID drawings U_t, each from the above distribution for given p. Let N be $\sum_{t=1}^{n} Y_t$. The value of N is thus the number of drawings with value $(1 - p)/p^2$. We see that $\Pr(N = 0) = (1 - p)^n$. The t statistic for a test of the hypothesis that $\mathrm{E}\, U = 0$ can be written as

$$T = \frac{\hat{\mu}}{\hat{\sigma}_\mu}, \text{ where } \hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} U_t, \text{ and } \hat{\sigma}_\mu^2 = \frac{1}{n(n-1)}\sum_{t=1}^{n}(U_t - \hat{\mu})^2.$$

Conditional on $N = 0$, $\hat{\mu} = -1/p + \bar{W}$, where $\bar{W} = n^{-1}\sum W_t$ is the mean of the W_t. Thus the conditional distribution of $n^{1/2}\hat{\mu}$ is $\mathrm{N}(-n^{1/2}/p, 1)$. Then, since $U_t - \hat{\mu} = W_t - \bar{W}$ if $N = 0$, the conditional distribution of $n\hat{\sigma}_\mu^2$ is $\chi^2_{n-1}/(n - 1)$. Consequently, the conditional distribution of T is noncentral t_{n-1}, with noncentrality parameter $-n^{1/2}/p$. We can compute as follows for $c > 0$:

$$\Pr(|T| > c) > \Pr(T < -c) > \Pr(T < -c \text{ and } N = 0) = \Pr(N = 0)\Pr(T < -c \,|\, N = 0).$$

Now

$$\Pr(T < -c \,|\, N = 0) = F_{n-1,-n^{1/2}/p}(-c),$$

where $F_{n-1,-n^{1/2}/p}$ is the CDF of noncentral t with $n - 1$ degrees of freedom and noncentrality parameter $-n^{1/2}/p$.

For fixed c and n, let $p \to 0$. Then we see that $\Pr(N = 0) \to 1$. Further, it is clear that $\Pr(T < -c \,|\, N = 0)$ also tends to 1, since the noncentrality parameter tends to $-\infty$, which means that the probability mass to the left of any fixed value tends to 1. It follows from this that the rejection probability tends to 1 whatever the critical value c, and so the test has size 1 if DGPs characterised by random variables distributed according to the present scheme are admitted to the null hypothesis.

Whatever restriction we may impose in order to restore the possibility of valid inference, we know that the restriction must exclude the distributions with small p. We are therefore led to consider a *uniform* bound on some higher moment. We want to show that such a

bound renders the mapping from $\mathcal{F}$ to the expectation continuous. Suppose then, that $\mathcal{F}$ is restricted so as to contain only distributions such that, for some $\theta > 0$, $\mathrm{E}|U|^{1+\theta} < K$, for some specified K.

A short proof shows that this restriction restores continuity to the mapping from the model $\mathbb{M}$ to the expectation of the distribution F. Note, however, that, if we want to know the size of a test concerning the value of the expectation, it is not enough to suppose the existence of the bound K. We must actually specify what K is numerically in order to bound the Type I error of the test.

The Bayesian approach normally does not come up against the Bahadur-Savage problem, because the likelihood is parametrically specified. Of course, the model containing the troublesome distributions above is also parametrically specified. Here, then, is the first of my questions that I wish to pose to Bayesians:

Question:

If we wish to be as agnostic about distributions as do the nonparametric frequentists, within the Bahadur-Savage limits, is there a Bayesian procedure that can attain comparable generality? In particular, do the Bahadur-Savage limits manifest themselves in the Bayesian approach?

4. Estimating Equations

The **generalised method of moments (GMM)** has had much success with econometricians since it was introduced by Lars Hansen in Hansen (1982). An embarrassingly large number of econometricians don't seem to know of the concepts of **estimating functions** and **estimating equations**, introduced by V. P. Godambe back in 1960; see Godambe (1960) and Godambe and Thompson (1978). It took a while for people to notice, but GMM and the methods related to estimating equations are more or less the same thing.

A starting point for understanding how to build models based on estimating functions is the notion of an **elementary zero function**. A function $f_t(y_t, \boldsymbol{\theta})$ is an elementary zero function associated with the observation y_t of a sample if the expectation of $f_t(y_t, \boldsymbol{\theta}_0)$ is zero, where $\boldsymbol{\theta}_0$ is the true parameter vector (very non-Bayesian!). A little more formally, if we start from a model $\mathbb{M}$ and a parameter-defining mapping $\boldsymbol{\theta} : \mathbb{M} \to \Theta$, then a function f is a zero function for this model if

$$\mathrm{E}_\mu\big(f(\boldsymbol{y}, \boldsymbol{\theta}(\mu))\big) = 0 \quad \text{for all } \mu \in \mathbb{M}.$$

This states that, if the zero function is evaluated at the true parameters of some DGP, then its expectation under that DGP is zero. To be useful, a zero function should have a nonzero expectation when evaluated at a value of $\boldsymbol{\theta}$ other than the true value. An elementary zero function is just a zero function that is associated with a specific observation in the sample.

Alternatively, given a set of elementary zero functions, we could define the model $\mathbb{M}$ so as to contain all DGPs for which the functions are indeed zero functions. Here, of course,

we run head on into the Bahadur-Savage problem. Since all that we are specifying is an expectation (read "moment"), inference is impossible until we impose further restrictions.

If we are prepared to impose such restrictions, then, as the GMM people have abundantly demonstrated, we can set up and estimate models of formidable generality. Thus:

Question:

What Bayesian approach can allow us to base models on specified moments?

This question is quite probing, since the notion of an elementary zero function is a very profound one. In fact, an elementary zero function can be considered the fundamental unit of statistical information. It is informative (or not) about the model parameters, and, being associated with one single observation, it lets us know just how informative that observation is about the various model parameters.

One can, theoretically at least, define the **optimal instrument** for a particular parameter and a particular elementary zero function. The definition is as follows:

$$F_{ti}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} \left(\frac{\partial f_t}{\partial \theta_i} \,\middle|\, \mathcal{F}_t \right),$$

where the set of sigma-algebras $\mathcal{F}_t$ constitutes a filtration that specifies the information available for conditioning with observation t. Formally,

$$\mathrm{E}_{\mu}\big(f_t(y_t, \boldsymbol{\theta}(\mu)) \,|\, \mathcal{F}_t\big) = 0.$$

If we can arrange things so that the zero functions f_t are homoskedastic and serially uncorrelated, then the optimal estimating equation for θ_i is

$$\sum_{t=1}^{n} F_{ti}(\hat{\boldsymbol{\theta}}) f_t(y_t, \hat{\boldsymbol{\theta}}) = 0.$$

In fact, the optimal instrument $F_{ti}(\boldsymbol{\theta})$ is the weight given to the elementary zero function $f_t(y_t, \boldsymbol{\theta})$ in the estimating equation for θ_i that minimises the asymptotic variance of the estimator. Thus the optimal instrument can serve as a measure of the quantity of information about a particular parameter provided by a particular zero function. An optimal instrument $F_{ti}(\boldsymbol{\theta})$ that is zero indicates that the zero function $f_t(y_t, \boldsymbol{\theta})$ is uninformative about the parameter θ_i.

5. The bootstrap

The bootstrap has been among my most pressing research preoccupations for nearly fifteen years now. It seems to be one of the most promising statistical tools for inference. Unfortunately, many people think that the bootstrap is synonymous with resampling (à la Efron (1979)), but, nowadays at least, the bootstrap is by no means limited to methods based on resampling.

Let me state here two "Golden Rules" for bootstrapping. They are extensions and reformulations of "guidelines" for bootstrap hypothesis testing found in Hall and Wilson (1991).

Golden Rule 1:

> The bootstrap DGP μ^* must belong to the model $\mathbb{M}$ that represents the null hypothesis.

If, in violation of this rule, the null hypothesis tested by the bootstrap statistics is not satisfied by the bootstrap DGP, a bootstrap test can be wholly lacking in power. Test power springs from the fact that a statistic has different distributions under the null and the alternative. Bootstrapping under the alternative confuses these different distributions, and so leads to completely unreliable inference, manifested by extreme loss of power, even in the asymptotic limit.

Golden Rule 2:

> Unless the test statistic is pivotal for the null model $\mathbb{M}$, the bootstrap DGP should be as good an estimate of the true DGP as possible, under the assumption that the true DGP belongs to $\mathbb{M}$.

A **pivotal** quantity for a given model is a function of the data generated by a DGP in the model the distribution of which is the same for all DGPs in the model. How this second rule can be followed depends very much on the particular test being performed, but quite generally it means that we want the bootstrap DGP to be based on estimates that are *efficient* under the null hypothesis.

Question:

What is the Bayesian bootstrap, and for what is it useful?

I ask this question largely out of ignorance, since there have been few Bayesian contributions to the bootstrap literature. One early article, Rubin (1981), is entitled "The Bayesian Bootstrap", and, as only to be expected at that time, equates the bootstrap with resampling. What is plainly a more sophisticated, and more distinctly Bayesian, approach is taken by Lo (1987). Alas, it also makes use of asymptotics.

Let me take this opportunity to denounce a view that is too often heard, namely that the bootstrap is an asymptotic procedure. It is not. What is true is that *current bootstrap theory* is very heavily dependent on asymptotic notions, but that is a failing of the theory, not the bootstrap.

Although students and others sometimes find their heads spinning on first acquaintance with the bootstrap, it is in fact a simple enough idea, muddled and confused by the term "bootstrap", with its unreasonable suggestion of getting something for nothing. But its basis is unrepentantly frequentist. The "true DGP" is essential to it, and inference is based on bootstrap P values that estimate a rejection probability under that true DGP.

My meagre reading on the Bayesian bootstrap leads me to believe that it is a resampling procedure. Of course, it is intended to estimate by simulation a posterior distribution,

something that a great many Bayesian computational methods try to do. But that is, I believe, almost incidental to the use of the bootstrap, which, after all, can be put to very numerous uses.

I think it is fair to say that the Bayesian approach is more parametrically oriented than most currently fashionable frequentist approaches. The very best bootstrap, in terms of reliability of inference, and also of theoretical tractability, is the parametric bootstrap, which, although simulation-based, makes no use of resampling.

Question:

Would it not be interesting to see how to apply the parametric bootstrap to Bayesian inference?

I can almost see in my head how to go about it! I would be more convinced by that than by any resampling Bayesian bootstrap.

6. Simulation

There is no doubt that the entire Bayesian enterprise has been "rebooted" by the rediscovery of the Gibbs sampler, and the other MCMC methods, based on the Metropolis-Hastings algorithm. I remember asking Tony Lancaster why the Gibbs sampler is so called, and being told (quite correctly) that it is in honour of J. Willard Gibbs, the great American father of statistical physics. However, "in honour of" does not mean "invented by". The Metropolis-Hastings crew, however, were also physicists (including the dreaded Edward Teller), who, at the very outset of the computer era, developed an algorithm for simulating equations of state based on the Boltzmann distribution.

Question:

Just how reliable are MCMC methods?

The idea is of course beautifully elegant. And it certainly suits the Bayesian program, so full of conditional distributions. But the conditions for it to work are rather restrictive. The expressions "ergodicity" and "metrical transitivity" come back to me from the time of my work in statistical physics, and these issues were regarded with great caution, as it was known that these conditions cannot be taken for granted.

It's not just a question of "burn-in", as it might be with the simulation of a time series of which the stationary distribution is unknown, as is still the case with ARCH and GARCH models. (The alphabet soup stands for **autoregressive conditional heteroskedasticity**, and the same thing **generalised**.) If that was all, then we could just let the computer run as long as needed. But, even once one believes that a simulated value is a realisation of the desired distribution, there is little guarantee that the whole sequence of subsequent realisations can be considered a random sample from that distribution, since the algorithm can be imprisoned – not for ever, just a long time – in a region of the distribution from which escape is a low-probability event.

Question:

Have there been improvements I haven't heard about in this technology?

Question:

What about the curse of dimensionality? Does it affect MCMC simulations?

7. Nonparametrics

In the frequentist approach to nonparametric problems, nonparametric regression for instance, a parametric specification of the DGPs in the model under study is replaced by conditions on the smoothness of otherwise unknown functions. I confess to a bit of a dislike for the theory of most of these methods, since it relies on asymptotic theory of a particularly ugly kind. The "asymptotic fictions" (I quote my co-author James MacKinnon in using this phrase) needed to prove some results are not only complicated and hard to interpret, but also provide no usable guide for such practical issues as the choice of bandwidth.

Question:

Is it possible to formulate smoothness conditions in the context of a prior distribution?

Since frequentists deal with function spaces here, the spectre of the non-existence of Haar measure raises its head again, since these spaces are not locally compact. Is that a genuine obstacle? Or perhaps Bayesians have a quite different approach (that I don't know about) to what the frequentists think of as nonparametric problems.

8. Some other topical issues

In this section, I mention some topics that have generated much interest and voluminous literature in the last couple of decades. My questions are all motivated by a wish to know how Bayesians approach these topics, and whether they have insights so far denied to frequentists.

Weak identification

For the last dozen years or so, the frequentist literature has betrayed an obsession with the problem of "weak instruments". More recently, it has been seen that it is more profitable to think of the problem as one of weak identification, since there are circumstances in which instrumental variables are not directly involved that also display the same sorts of difficulties for estimation and inference as the weak-instrument problems. The latter were of course inspired by empirical work, and so were of immediate practical interest. But other problems characterised by weak identification are not devoid of interest.

The work on weak instruments has brought back into fashion work of the 1970s and 1980s on the exact distributions of IV estimators, always under the assumption of Gaussian disturbances. That work, frequentist to the core, had been more or less set aside once GMM came in. It has, however, succeeded in clearing up some problems revealed by the weak instrument literature, especially when the bootstrap is used intelligently for implementing robust procedures, as in Davidson and MacKinnon (2008) and (2010).

Question:

Do Bayesians regard these problems as non-problems that disappear when priors and likelihoods are properly specified?

It is certainly true that combining prior information with sample information can alleviate identification issues, but frequentists would dispute the existence of any relevant prior information in most cases they consider.

If Bayesians think that the Bayesian approach really does sidestep issues of weak identification, then what do they think of this very voluminous literature? One answer might be to say that it was valuable to be shown empirically that some earlier conclusions based on estimation with what turned out to be weak instruments were unfounded, while still maintaining that all the theoretical follow-up on weak instruments is sound and fury. I would be much interested to hear a considered Bayesian response to this question.

Distribution-Free Inference

Much statistical and econometric work on income distribution is said to be "distribution free". This means in practice that we can estimate the standard error of estimates of quantities of interest, like the Gini index or other indices of inequality or poverty, without making assumptions on the functional form of the distribution from which our sample was drawn.

This is of special value for bootstrapping, since it permits **pre-pivoting**, that is, the construction of test statistics that, under the null, have an asymptotic distribution that is nuisance-parameter-free.

Question:

Suppose I want to construct a confidence interval for an inequality index on the basis of a random sample from the population under study. How best to do this as a Bayesian?

In a way, I hope that there is no glib answer to this question. Just last year, I worked out a way to do asymptotic inference on the Gini index that turned out to give pretty reliable inference (using confidence intervals) when bootstrapped; see Davidson (2009). Can theoretical advances of this sort be put to constructive use in the context of Bayesian inference? The frequentist problem is characterised by being nonparametric – the distribution-free aspect – but concerned with a particular parameter, namely the index in question. It would seem hard to formulate a likelihood adapted to such a problem.

Unit roots and Cointegration

This is a topic where Bayesians may well have the right to call frequentist practitioners of the arcane art of working with cointegration misguided. In particular, I have some difficulty with the current fad for fractional cointegration, even though some of my best friends work on the topic.

At the 1995 World Congress of the Econometric Society, Chris Sims voiced some skepticism about the whole unit root enterprise. At the time, I didn't really agree with him, and I

still don't. But I do share his skepticism when it comes to fractional cointegration. I seriously doubt whether the economic time series we have at our disposal contain enough information to discriminate among the various models, including those that use fractional cointegration, that have been proposed to take account of long memory. In the keynote paper I gave at Rimini, I discuss unit root testing without actually advocating its use. My excuse is that the paper is not really about unit roots at all. They simply provide a convenient tractable example to illustrate some issues that pertain to the bootstrap.

Question:

Since Pierre Perron's pioneering work, the unit root business has been bound up with the thorny problem of structural change. How does a Bayesian approach this thorny problem?

Frequentists have a lot of trouble in formulating models that may or may not incorporate structural change, either at known dates or unknown dates. I recently heard Perron give a talk in which he makes assumptions about the *distribution* of structural breaks, and succeeds in fitting certain very long time series quite well. The work presented evolved into Perron and Qu (2007). Something like this would seem to correspond well with the Bayesian notion of a prior distribution. Is this something worth pursuing? (If so, good, since frequentists have got themselves into a big muddle over this business.) Incidentally, I see from Pierre Perron's website that he has recently been using Bayesian methods to further his studies of structural breaks.

Partial Specification

In my work with heavy-tailed distributions (commonly encountered with income data), it has sometimes proved profitable to adopt a basically nonparametric attitude, but with parametric modelling of the tail, or tails, of the distribution. This gives rise to a reasonably good bootstrap procedure – see Davidson and Flachaire (2007).

As is usual with nonparametric modelling, there are various tuning parameters that must be chosen, and, with luck, chosen in not too suboptimal a manner. Where does the nonparametric part stop and the parametric tail begin? How much of the sample should be used in order to estimate the parameters of the tail?

Question:

Is there any way to cope with a partial specification of a distribution of this sort in the Bayesian approach?

Roughly speaking, we want a parametric prior for the tail, and an uninformative prior for the main body of the distribution. I can think of a couple of ways I might set about trying to formulate a prior of this sort, but a real Bayesian would surely do better.

If this were to work satisfactorily, it might be possible to set up a frequentist bootstrap based on Bayesian estimation of a data set. In fact, there might be abundant opportunities to do something like that in different contexts. Recall my second Golden Rule of bootstrapping: we want as good an estimate as possible of the DGP. In cases in which one

might be persuaded that a Bayesian posterior distribution is as good an estimate as any other, maybe it would work well with the bootstrap.

Biostatistics

In my naive frequentist manner, I might have thought that biostatistics and epidemiology would provide an almost ideal area for the application of Bayesian methods, since I have always tended to think that a Bayesian approach is good for statistical decision making, while a frequentist approach is better suited for analysis of data. (I know that Bayesians contest this notion hotly, but that's why we're having this debate!)

Although I see evidence from Google that Bayesian methods are occasionally used in biostatistics, they don't seem to be the most popular choice. I would guess that there is abundant scope for technical improvements, from both sides of our divide, in this field, which is of considerable importance to the human race, and seems to attract a lot of research money.

Question:

What are the potentially most profitable areas in which frequentist and Bayesian statisticians could contribute to better technical skills in biostatistics?

9. Concluding remarks

All statisticians, whether Bayesian or not, share at least one goal, namely the efficient extraction of information from data sets. We may well differ, and will probably continue to differ, on how we wish to present and interpret the information so extracted, but, as an agnostic, I see no reason to adopt any attitude other than that people's preferences differ, as do the questions they ask, and so it is likely that some questions are best answered by imposing randomness on a parameter space, while others benefit from supposing a random sample space.

If one takes seriously the idea that an elementary zero function constitutes the fundamental unit of statistical information, then there could well be scope for a joint effort from Bayesians and frequentists in order to clarify statistical theory by starting from the concept of a set of elementary zero functions, and to develop improved methods of inference – on both sides of the fence.

Lastly, looking at things from both a frequentist and Bayesian perspective simultaneously – or almost so – leads to a notion that tidies up the oft-cited duality between hypothesis testing and the construction of confidence sets. A confidence set can be defined as the set of parameter values for which a hypothesis that those are the true parameters is not rejected at a significance level that is the complement of the desired confidence level. The set thus corresponds to a whole family of hypothesis tests. Conversely, a hypothesis test corresponds to a whole family of confidence sets, with the P value as the complement of the confidence level for which the hypothesised parameter vector lies on the frontier of the confidence set. If one were to plot a **confidence profile**, which for a scalar parameter would be a graph of the marginal confidence level as a function of the parameter value, then this

would be a frequentist construction that would have an interpretation not too different from that of a posterior density. It would not be the same thing as a posterior density at all, of course, but its uses would be very similar. It may be interesting to speculate whether there exists a duality relation between a confidence profile and a posterior density.

References

Bahadur, R. R. and L. J. Savage (1956). "The nonexistence of certain statistical procedures in nonparametric problems", *Annals of Statistics*, **27**, 1115–22.

Davidson, Russell (2009). "Reliable Inference for the Gini Index", *Journal of Econometrics* **150**, 30–40.

Davidson, Russell and E. Flachaire (2007). "Asymptotic and Bootstrap Inference for Inequality and Poverty Measures", *Journal of Econometrics* **141**, 141–66.

Davidson, Russell and James G. MacKinnon (2008). "Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables", *Econometrics Journal* **11**, 443–77.

Davidson, Russell and James G. MacKinnon (2010). "Wild bootstrap tests for IV regression", *Journal of Business and Economic Statistics*, **28**, 128–144.

Dufour, Jean-Marie (2003). "Identification, weak instruments, and statistical inference in econometrics", *Canadian Journal of Economics* **36(4)**, 767–808.

Efron, Bradley (1979). "Bootstrapping methods: Another look at the jackknife", *Annals of Statistics* **7**, 1–26.

Gelman, Andrew (2008). "Objections to Bayesian statistics", *Bayesian Analysis* **3(3)**, 445-450.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation", *Annals of Mathematical Statistics* **31**, 1208–11.

Godambe, V. P., and M. E. Thompson (1978). "Some aspects of the theory of estimating equations", *Journal of Statistical Planning and Inference*, **2**, 95–104.

Hall, Peter, and Susan R. Wilson (1991). "Two guidelines for bootstrap hypothesis testing", *Biometrics* **47**, 757–62.

Halmos, Paul (1950). *Measure Theory*, D. van Nostrand and Co., Section 52.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators," *Econometrica*, **50**, 1029–54.

Lo, Albert Y. (1987). "A large sample study of the Bayesian bootstrap", *Annals of Statistics* **15**, 360–75.

Perron, Pierre, and Zhongjun Qu (2007). "An analytical evaluation of the log-periodogram estimate in the presence of level shifts", Working paper.

Rubin, Donald B. (1981). "The Bayesian bootstrap", *Annals of Statistics* **9**, 130–4.

Rubinstein-Salzedo, Simon (2004). "On the existence and uniqueness of invariant measures on locally compact groups", working paper.

Senn, Stephen (2008). "Comment on Article by Gelman", *Bayesian Analysis* **3(3)**, 459-462.